Deep Hourglass for Brain Tumor Segmentation

Eze Benson[1], Michael P. Pound[1], Andrew P. French[1,2], Aaron S. Jackson[1], Tony P. Pridmore[1]

[1] School of Computer Science, University of Nottingham
[2] School of Biosciences, University of Nottingham
ezenwoko.benson@nottingham.ac.uk
michael.pound@nottingham.ac.uk
andrew.p.french@nottingham.ac.uk
aaron.jackson@nottingham.ac.uk
tony.pridmore@nottingham.ac.uk

Abstract. The segmentation of a brain tumour in an MRI scan is a challenging task, in this paper we present our results for this problem via the BraTS 2018 challenge, consisting of 210 HGG and 75 LGG volumes for training. We train and evaluate a CNN encoder-decoder network based on a singular hourglass structure. The hourglass network is able to classify the whole tumour (WT), enhancing (ET) tumour and core tumour (TC) in one pass. We apply a small amount of preprocessing to the data before feeding it to the network but no post processing. We apply our method to two different unseen sets of volumes containing 66 and 191 volumes. We achieve an overall Dice coefficient of 92% on the training set. On the first unseen set our network achieves Dice coefficients of 0.66, 0.82 and 0.72 for ET, WT and TC. On the second unseen set our network achieves Dice coefficients of 0.62, 0.79 and 0.65 on ET, WT and TC.

Keywords: Convolutional Neural Network, Deep Learning, Hourglass, Glioma

1 Introduction

Identifying regions of the brain which are tumourous is a task often carried out by medical professionals. Manually classifying segments of the tumour is a subset of a group of problems commonly referred to as semantic segmentation. Semantic segmentation is the task of assigning a class to each pixel within an image, modern automated solutions to this problem often use convolutional neural networks (CNN). The introduction of fully convolutional networks (FCN) [1] established a convolutional neural network architecture that is widely used for the task of semantic segmentation. Architectures such as U-NET [2] achieved success in biomedical imaging by adopting a similar architecture.

We propose the use of an adapted hourglass [3] network to solve the problem of tumour segmentation. The hourglass network improves on U-NET by using bottleneck blocks and adding convolutions to the skip connections. Training a CNN for this

problem is a natural choice as they have demonstrated state-of-the-art performance on semantic segmentation problems such as the widely used Pascal VOC2012 [9] and cityscapes [10] datasets.

2 Methods

2.1 Data

The dataset of BraTS 2018 [4-7] provides defined training and validation sets. The training set is composed of 210 MRI scans of high grade gliomas (HGG) and 75 MRI scans of low grade gliomas (LGG). Whilst the validation set is a group of 66 mixed HGG and LGG tumours. The MRIs are volumes with XxYxZ dimensions of 240x240x155. Each volume has four corresponding modalities FLAIR T1,T2 and T1CE.

2.2 PreProcessing

A high variance in intensity in both validation and training set was observed this lead us normalise the training set to be centred around zero with a standard deviation of one. By normalizing the data, we found that the required training time was reduced and the accuracy of the network was increased. The formula for normalization is given in figure 1. Each modality was normalized separately due to the variance in intensity profile between modalities.

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

Where x is the current intensity, μ is the mean of the modality and σ is the standard deviation of the modality.

2.3 Hourglass Architecture

Our approach is to handle 2D slices of each volume separately, a 2D semantic segmentation problem. We performed additional experimentation using a volumetric encoder-decoder but found that the benefit of an end-to-end volumetric approach was outweighed by the significant necessary drop in features at each layer due to memory restrictions.

We design our network using an encoder-decoder structure, adapted from an hourglass network, popularized in the domain of human-pose estimation [3] The structure of the hourglass is similar to other encoder-decoder networks, but contains a denser use of residual blocks throughout.

The encoder contains 7 residual bottleneck blocks [8], after each a max-pooling layer performs spatial downsampling. A further three residual blocks at the lowest spatial resolution derive higher-level features before a series of bilinear upsampling operations return the network to the original spatial resolution. As in the encoder, all upsampling

operations of the decoder are interleaved with residual blocks. Skip layers are added between each matching resolution of the encoder and decoder, with each containing an additional residual block to learn an appropriate mapping.

In order to improve the network's results for the final test set we made architectural changes to improve accuracy whilst keeping memory consumption to a minimum. We found that the choice of upsampling layer (e.g. bilinear, max-unpooling [11]) made little difference to the performance of the network. Unlike the original work [3] we chose not to stack hourglass networks sequentially and perfom intermediate supervision, we found this too had a negligible effect on performance. The number of spatial-downsampling layers, 7 in total, were originally chosen based on the input resolution. However, through experimentation we found that using 5 downsampling layers was optimal and save memory. Only one residual block is used at each depth because adding two at all depths immediately doubles memory consumption which surpasses current memory constraints. We also found that replacing elementwise summation with concatenation followed by a 1x1 convolution improved results noticeably. Despite the additional memory consumption of the concatenation and convolutional layer, the increase in performance boost makes the change worthwhile.

2.4 Training

The training was split into two phases pre and post true validation set release. In the first phase the dataset was split into a test set, validation set and training set where each set was 10%,10% and 80% of the original training set respectively. The data provided is treated as though it is the entire dataset so that our training can be validated and tested in preparation for the true validation set. This allows the network to avoid overfitting and approximate the results expected on the release of the second dataset. Later the network is retrained using a 10% test set and 90% training set split in order to obtain test results on the original data whilst maximizing the training set size. The network is trained for the same number of epochs for all training. The second phase is conducted post true validation set release. In this phase the BraTS dataset is split into 10% validation and 90% training.

The network is trained using an identical training scheme for both the natural and augmented dataset.

The hourglass network implemented in this paper only uses spatial convolutions which means the data must be sliced along the depth dimension which effectively converts a volume into a series of 155 image with a spatial resolution of 256^2. The volumes have a spatial resolution of 240^2 however for convenience we pad them to the new resolution. The dataset used is then 44175 images instead of 285 volumes. All four modalities are used to train the network and are inputted as different channels to the network.

The hourglass is trained using a cross entropy loss function with a learning rate of 10^{-5} which is decreased by a factor of 10 every 30 epochs. The network is trained for a total of 50 epochs therefore the learning rate is only adapted once. The adaptive gradient descent algorithm, RMSProp is used to train the network faster than the typical stochastic gradient descent.

2.5 Data Augmentation

Two methods of data augmentation are used in this paper vertical flipping and random intensity variation. Vertical flipping is used because it matches the natural symmetrical shape of the brain.

Random intensity variation is used because the intensity between MRI scans varies significantly. This is shown by the fact that the standard deviation of the FLAIR modality in the dataset is greater than the mean by almost a factor of 10. E.g. The standard deviation and mean for the FLAIR modality are 529.2 and 61.8 respectively. The T1, T1CE and T2 modalities have similar standard deviations. Intensity variation is performed on the normalised dataset by first rescaling the standard deviation of the dataset and then shifting the mean. This allows the dataset to include image intensities which are not present in the original dataset but could appear on an MRI volume. The range for randomly changing the standard deviation is between zero and two. The mean is shifted between values of 0.4 and -0.4. Values above a standard deviation of two were experimented with but lead to a significant decrease in accuracy. Shifting the mean by over 0.5 and under -0.5 were trialed but also caused an accuracy decrease.

3 Results & Discussion

The results are split into three sections, the results on the training data set, the results on the later released validation set and the results on the final test set. Results are shown for networks trained on the standard data and on augmented data in the validation set.

3.1 Training Dataset

We trained the network on 90% of the data leaving 10% for testing purposes. The network achieved a Dice coefficient of 92% with an IOU of 86%. We find that IOU approximates the network's worst performance on the test set in contrast to Dice which gives an approximate representation of the average case.

3.2 Validation Dataset

The results presented in this section are those achieved when segmenting the validation set using the network trained in section 3.1. Table 1 shows the results of the segmentation without augmentation and table 2 shows the results with flipping and intensity variation. The metrics provided in both tables are the standard metrics output by the BraTS automatic online evaluation server. Some metrics have been omitted to save space, only the most important evaluation metrics have been included.

Table 1. The results of the hourglass network segmenting the unseen validation set without augmentation in the training data.

	Dice ET	**Dice WT**	**Dice TC**	**Hausdorff ET**	**Hausdorff WT**	**Hausedorff TC**
Mean	0.59102	0.81638	0.63479	18.11974	94.28005	130.6982
Std	0.28441	0.12274	0.24233	26.62022	50.15014	42.39722
Median	0.70712	0.86227	0.70526	5.73233	97.12933	132.5891
25 Quantile	0.47849	0.7832	0.50762	3.16228	52.71734	103.3565
75 Quantile	0.80193	0.89504	0.82568	20.02904	135.8139	163.3901

Table 2. The results of the hourglass network segmenting unseen validation set where the network has been trained with augmented data

	Dice ET	**Dice WT**	**Dice TC**	**Hausdorff ET**	**Hausdorff WT**	**Hausedorff TC**
Mean	0.56337	0.82204	0.61797	14.27762	13.57432	17.94668
Std	0.2888	0.12962	0.21713	23.25875	15.31909	18.13535
Median	0.66684	0.86688	0.6712	5.91548	6.59447	11.18034
25 Quantile	0.39652	0.78219	0.49556	2.82843	4.18205	8.29669
75 Quantile	0.79578	0.90147	0.79259	12.55482	14.96802	18.78738

After comparing the metrics between a dataset with augmentation and one without we find that in this challenge augmentation appears to give a small increase in accuracy for Dice coefficient and improves the Hausdorff accuracies significantly. It is likely the case that the frequency at which the network misclassifies pixels remains similar but the network's ability to localize the pixels is increased.

Overall the network segments the whole tumour more accurately than it does the core tumour or enhancing tumour, from the results in previous challenges this result is expected. Naturally the enhancing and core tumour are much more difficult to segment due to the similarity between all classes.

Table 1 and table 2 both show a large disparity between the median and mean accuracy especially with results for the enhancing tumour where the difference is around 10%. The difference is caused by the difficulty of detecting the enhancing tumour and core tumour in some volumes. In most volumes the Dice coefficients are well above the mean however some outliers achieve a score of 0 therefore reducing the mean significantly. When removing these cases the mean Dice coefficient increases by 4% showing that the disparity can be explained by a few very difficult volumes. Some examples of the metrics achieved on these volumes are shown in table 3.

Table 3. Segmentation results for very difficult volumes using a network trained with augmented data

	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausedorff TC
Brats18_TCIA09_248_1	0	0.79274	0.62938	0	14.17745	10.81665
Brats18_TCIA10_195_1	0	0.79941	0.62923	0	15.23155	25.98076
Brats18_TCIA11_612_1	0	0.73506	0.60061	0	52.7731	48.51804
Brats18_TCIA12_613_1	0	0.69355	0.2629	0	49.96999	9
Brats18_TCIA13_646_1	0	0.90187	0.3996	0	35.82736	6.48074

3.3 Test Dataset

Before the release of the final evaluation dataset we train our network using 95% of the training data. The remaining 5% of the training data is used for on the fly validation of the network to monitor training and prevent overfitting. The network architecture has been adapted to improve the results on the validation set, these architectural changes are discussed in section 2.3. We present the new validation set results along with the test set results. Section 3.2 showed that the network has an increase in Hausdorff95 accuracy when data augmentation was used. The network used for the results in this section was trained using data augmentation.

Table 4. The results of the hourglass network segmenting unseen the validation set

	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausdorff TC
Mean	0.65968	0.82035	0.72126	15.93993	26.41346	18.86627
Std	0.2761	0.0951	0.22999	25.56283	23.61084	20.5594
Median	0.78917	0.84303	0.79516	4.68556	17.32471	12.46982
25quantile	0.56348	0.77912	0.61818	2.44949	7.19076	6.61366
75quantile	0.84395	0.88679	0.88725	17.60682	38.13037	19.60737

Table 5. The results of the hourglass network segmenting unseen test set

Label	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausdorff TC
Mean	0.616999	0.786054	0.651771	47.48392	13.53952	31.58409
Std	0.323864	0.248256	0.336683	113.7593	23.50906	83.237
Median	0.77046	0.87744	0.82395	3	5	6.40312
25quantile	0.473615	0.7986	0.475805	1.73205	3	3.31662
75quantile	0.846715	0.920085	0.895065	9.842435	9.72086	14.7985

Table 4 shows the results of the hourglass network on the validation set. The dice scores for the validation set increase by 10% for both ET and TC whilst remaining approximately the same for the whole tumour segmentation. Conversely the Hausdorff scores increase (where a higher score is a decrease in performance) by 1, 13 and 2 for ET,WT and TC respectively. The increase in dice score indicates that the total number of pixels that are being classified correctly has increased but the decrease in Hausdorff score shows that the largest error in the shape of the classified pixels is much higher. The qualitative analysis presented in section 3.4 shows that this may be because misclassification of background pixels far away from the site of the tumour.

The median Hausdorff distance and dice score are significantly better than the mean indicating that the mean results are being distorted by a small subset of difficult to segment brain tumour volumes. This is discussed in section 3.2. The std of both metrics is also very high showing that the networks performance varies largely between volumes. The network shows a significant improvement in the most problematic volumes highlighted in table 3. Tables 6 shows the modified network's performance on the selected examples. The average Hausdorff distance for the selected examples indicates an overall performance decrease however performance on individual volumes varies significantly when dice scores are compared. The network architecture was modified in order to increase performance on the enhanced tumour, table 6 shows that on 3 out of 5 selected cases there is an increase of between 4.6% and 38% for the enhancing tumour dice score. The variability in dice score amongst the other two metrics indicates that the training scheme has altered the networks ability to classify the tumour in these volumes.

Table 6. The modified network's segmentation results on a subset of problematic volumes

	Dice ET	Dice WT	Dice TC	Hausdorff ET	Hausdorff WT	Hausedorff TC
Brats18_TCIA09_248_1	0	0.79768	0.4842	0	61.25561	12.40967
Brats18_TCIA10_195_1	0	0.85514	0.71116	0	22.67157	30.23243
Brats18_TCIA11_612_1	0.38484	0.63438	0.39605	98.47207	59.87487	98.2527
Brats18_TCIA12_613_1	0.05991	0.93749	0.93743	58.25805	4.12311	2.82843
Brats18_TCIA13_646_1	0.0459	0.69703	0.61365	111.1884	87.67696	15.13275

The test set results show a decrease in performance on both dice score and hausdorff distance when compared to validation set results. The median scores for both metrics are noticeably better. This indicates that the validation set contains easier to segment volumes but the ratio between difficult and easy volumes is higher. The test set appears to have much more difficult volumes, this is corroborated by the very high standard deviation values. The results suggest that the percentage of easily segmented volumes in the test set is higher than the validation set.

Despite the differences between the network's performance on the validation and test sets both Table 4 and 5 indicate the same overall strengths and weaknesses of the network as well as the difficulties within the dataset.

3.4 Qualitative analysis

In this section we present singular slices taken from the network output. The output has 4 classes which are represented by 4 different colours in the segmentation map. Black, yellow, blue and red represent background, whole tumour, core tumour and enhcnaing tumour.

The network makes many mistakes when segmenting unseen volumes, most often these errors are misclassifying healthy brain tissues as tumourous. Often the mistakes are of a small area which does not affect the dice score significantly but has a noticeable impact on the hausdorff distance. These errors are important and can be improved upon however for brevity this section will focus on the largest errors associated with the problematic volumes highlighted in section 3.3. Figure 1 shows large errors in classification. The largest errors the network makes occur when the input image has large errors of darkness within the tumour caused by necrosis or an irregular tumour shape. It is unclear why this occurs but could be because the training set contains mostly tumour which have small amounts of necrosis which are masses enveloped by the whole tumour. Therefore when given to the network it is unable to deal with the variance.

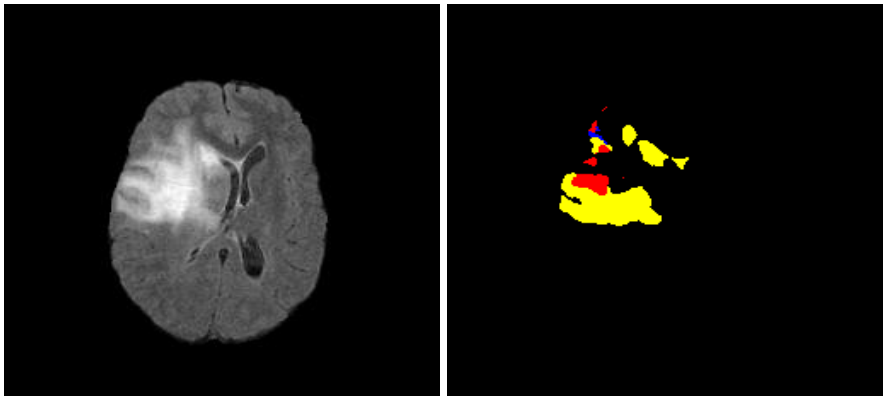

Figure 1*: Left, a FLAIR volume slice containing both brain and tumour tissues. Right, a slice from the network output showing erroneous segmentation results.*

4 Conclusion

We propose a solution which achieves a 92% Dice coefficient on the training set and 0.66, 0.82 and 0.72 on the validation set. On the test set the network achieves 0.62,0.79 and 0.65 Dice scores. Although the network underperforms on Dice score it can achieve a competitive Hausdorff distance.

Much of the network's underperformance is related to outliers in the set which could be mitigated in future with better preprocessing techniques. Future networks should train more on these difficult volumes using wider public datasets or through synthetic images generated by a CNN. Memory consumption is often a problem when using CNNs, to combat this we plan to add residual blocks in depths which increase the overall accuracy of the network the most. We also plan to add skip connections with an inception block structure [12] as shown in [13] to increase accuracy further.

We show that 2D architectures can segment 3D volumes with success but require fine tuning and a deeper architecture to achieve better results. An approach to bridge the gap may between 2D and 3D may be required. 3D networks outperform 2D networks when depth context is key how much context is required in most tasks remains unclear. We plan to use a 2.5D approach where each slice has an accompanying adjacent slice either side to provide some depth context.

5 References

1. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 3431-3440).
2. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InInternational Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234-241). Springer, Cham.
3. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. InEuropean Conference on Computer Vision 2016 Oct 8 (pp. 483-499). Springer, Cham.
4. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging. 2015 Oct;34(10):1993.
5. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific data. 2017 Sep 5;4:170117.
6. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby J, Freymann J, Farahani K, Davatzikos C. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive. 2017;286.
7. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby J, Freymann J, Farahani K, Davatzikos C. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive. 2017;286.
8. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).
9. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. International journal of computer vision. 2010 Jun 1;88(2):303-38.
10. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 3213-3223).
11. Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561. 2015 Nov 2.
12. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 1-9).
13. Bulat A, Tzimiropoulos G. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. InThe

IEEE International Conference on Computer Vision (ICCV) 2017 Oct 1 (Vol. 1, No. 2, p. 4).